Improving the quality of discrete-choice experiments in health: how can we assess validity and reliability?

Ellen M. Janssen, Deborah A. Marshall, A. Brett Hauber & John F. P. Bridges

To cite this article: Ellen M. Janssen, Deborah A. Marshall, A. Brett Hauber & John F. P. Bridges (2017) Improving the quality of discrete-choice experiments in health: how can we assess validity and reliability?, Expert Review of Pharmacoeconomics & Outcomes Research, 17:6, 531-542, DOI: 10.1080/14737167.2017.1389648

To link to this article: http://dx.doi.org/10.1080/14737167.2017.1389648
Improving the quality of discrete-choice experiments in health: how can we assess validity and reliability?

Ellen M. Janssen\textsuperscript{a}, Deborah A. Marshall\textsuperscript{b}, A. Brett Hauber\textsuperscript{c} and John F. P. Bridges\textsuperscript{a}

\textsuperscript{a}Department of Health Policy and Management, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA; \textsuperscript{b}Department of Community Health Sciences, University of Calgary, Health Research Innovation Centre (HRIC), Calgary, AB, Canada; \textsuperscript{c}Research Triangle, RTI Health Solutions, NC, USA

ABSTRACT

Introduction: The recent endorsement of discrete-choice experiments (DCEs) and other stated-preference methods by regulatory and health technology assessment (HTA) agencies has placed a greater focus on demonstrating the validity and reliability of preference results.

Areas covered: We present a practical overview of tests of validity and reliability that have been applied in the health DCE literature and explore other study qualities of DCEs. From the published literature, we identify a variety of methods to assess the validity and reliability of DCEs. We conceptualize these methods to create a conceptual model with four domains: measurement validity, measurement reliability, choice validity, and choice reliability. Each domain consists of three categories that can be assessed using one to four procedures (for a total of 24 tests). We present how these tests have been applied in the literature and direct readers to applications of these tests in the health DCE literature. Based on a stakeholder engagement exercise, we consider the importance of study characteristics beyond traditional concepts of validity and reliability.

Expert commentary: We discuss study design considerations to assess the validity and reliability of a DCE, consider limitations to the current application of tests, and discuss future work to consider the quality of DCEs in healthcare.

1. Introduction

In recent years, there has been a general trend toward patient involvement in treatment [1,2], research [3,4], and regulatory decision-making [5,6]. As regulatory and health technology assessment (HTA) decision-making [7,8] is becoming more patient centered, ways to measure preferences of healthcare stakeholders are being explored [9]. Patient preferences play an important role in many quantitative approaches to benefit-risk assessment [10], and stated-preference methods such as discrete-choice experiments (DCEs) are increasingly used to study acceptance of benefit-risk tradeoffs [11]. A growing number of HTA agencies are now formally incorporating patient preferences into their decision-making for regulatory approval, listing and reimbursement, and pricing [7,12,13], and the US FDA has already provided guidance on quantitative preference assessment for regulatory approval of medical devices [14].

Moving forward, preference studies will need to meet evidence standards consistent with standards for clinical evidence [9]. Guidance from the FDA on incorporating patient preference information in benefit-risk assessments calls for checks on the quality of stated-preference studies [14], including the logical soundness and validity of patient-preference results, but do not specify how to measure these concepts. Validating stated-preference assessments is difficult because of the hypothetical nature of the choice scenarios [15]. Whether participants would display the same preferences if they were presented with these choices and would subsequently experience the consequences of their choices, is unknown.

The validity of stated-preference methods when compared with revealed preferences has been explored in transport, marketing, and environmental economics [16,17]. Opportunities to observe preferences through real-life choices in healthcare are limited [11] because healthcare options might not be available on the market yet [14] or choices are made on behalf of the patient by a physician [18] or a third-party payer [19,20]. A recent systematic review in the application of DCEs in environmental economics found that environmental DCEs generally provide limited evidence on the internal reliability and validity of DCEs [21]. The Medical Device Innovation Consortium (MDIC), a public–private partnership to advance medical device regulatory science, has called for studies that evaluate specific aspects of the validity of DCEs that are applied to healthcare [9]. While literature reviews on the application of DCEs in healthcare have discussed tests of internal validity, the tests discussed in these reviews have not been comprehensive [22–24]. Furthermore, these discussions often do not provide details on how to conduct tests for validity and reliability. Last, it is not clear which study qualities, including but not limited to validity and reliability, should be prioritized when conducting a DCE.

CONTACT Ellen M. Janssen ejanssen@jhu.edu Department of Health Policy and Management, Johns Hopkins Bloomberg School of Public Health, 624 N Broadway, Baltimore, MD, USA

© 2017 Informa UK Limited, trading as Taylor & Francis Group
The MDIC report defines validity of a stated-preference study as ‘the extent to which quantitative measures of relative importance or tradeoffs reflect the true preferences of patients’ [9]. In our study, we expand on the MDIC definition to include both validity and reliability and apply this in the context of the healthcare DCE literature. We present a practical list of tests of validity and reliability that have been applied in the health DCE literature and describe how these tests have been conducted in published studies. We also consider what other desirable and actionable study characteristics, outside of tests for validity and reliability, influence the quality of a preference study. Last, we discuss concepts that should be considered when designing a DCE to allow researchers to assess validity and reliability and we provide an overview of future work that should be carried out to ensure the study quality of DCEs applied in health. This review can help researchers consider questions of study quality when conducting a DCE and will help them identify and apply tests of validity and reliability.

2. Targeted literature search

We conducted a targeted literature review to identify articles that discussed the validity and reliability of DCEs as applied in the healthcare preference literature. The search was meant not to identify every published DCE that assessed validity or reliability but rather to identify studies that discussed concepts of validity and reliability and how these could be tested. We searched PubMed Central for DCEs that used the words validity and/or reliability and were published between January 2000 and January 2017. We also conducted a hand search to identify studies that reviewed DCEs or that assessed the reliability or validity of a DCE but were omitted in the targeted search. The literature search described was admittedly limited by including only one database and narrow search terms. It is likely that some applicable articles were missed in this review, but we believe that we identified the majority of test of validity and reliability that have been utilized in the health DCE literature.

We included papers that gave an overview of DCEs as a stated-preference method, discussed methods issues related to conducting DCEs, and mentioned processes to assess validity or reliability of results. We excluded studies that did not discuss DCEs or that did not discuss tests that could be used to evaluate the validity or reliability of a DCE. We also excluded studies that were not related to healthcare or that only employed qualitative methods. From the included literature, we extracted tests of validity and reliability.

The PubMed search identified 139 studies, and the hand search identified an additional four articles. Title and abstract, full-text review, and the extraction of tests of validity and reliability were conducted by one reviewer. After title and abstract review, 109 studies were excluded. Tests were extracted from 13 studies that were reviews or made a methodological contribution in describing and applying a test. Four of these studies were literature reviews [22–25], three provided overviews on how to conduct stated-preference studies [26–28], and seven provided a methodological explanation of a particular test with an empirical illustration [28–34].

3. Types of tests of validity and reliability for DCEs

To conceptualize the tests of validity and reliability, principles of concept mapping were adopted to sort and group tests [35]. A total of 24 tests of validity and reliability were identified from the literature. Table 1 presents a summary of each test and provides a reference to an application of each test. An iterative and collaborative process was then used to group the different tests based on the concept of validity and reliability they measured that had been discussed in the literature. Through this process, 12 conceptual groups were identified, with each group containing between one and four tests of validity or reliability. The conceptual groups were sorted into concepts that measured validity and those that measured reliability. Based on this sorting, a further distinction was made between concepts that applied to most survey-based studies and those that applied only to preference-based studies.

The developed conceptual framework included four conceptual domains: measurement validity, measurement reliability, choice validity, and choice reliability (Figure 1). Measures of validity capture how accurately an instrument measures the outcome of interest, whereas measures of reliability capture how consistently it measures the outcome of interest. In this conceptual framework, measurement validity and reliability capture traditional measures of validity and reliability [49] that apply to most scientific studies and are not limited to stated-preference methods.

Choice validity and choice reliability still address issues of accuracy and consistency, but these concepts examine whether the results of the study adhere to assumptions specific to choice behavior [50]. These concepts do not apply to studies that do not elicit measures such as preferences or choices. The axioms of utility theory lay the groundwork for many of the assumptions on consumer choice behavior used in tests of validity and reliability of DCEs. Since a thorough discussion of utility theory is beyond the scope of this study, we direct readers interested in learning more about utility theory and DCEs to a paper published by Hougaard et al. (2012) [51].

4. Measurement validity

Measurement validity captures whether the results from a DCE meet face/content validity, convergent validity, and external validity. These concepts identify how accurately the instrument measures preferences and how generalizable these preferences are to other circumstances.

4.1. Face/content validity (four tests)

Face and content validity are largely assessed through qualitative methods. Face validity, also referred to as theoretical validity [22,23,25,27], assesses the extent to which the results are consistent with a priori expectations. We prefer the term face validity because theoretical validity can also refer to the axioms of welfare theory, which may or may not be an appropriate standard in evaluating medical decisions [28,33]. Content validity refers to the extent to which the DCE takes account of all things deemed important in the construct’s domain [25]. In a DCE, this refers to the attributes and levels...
Table 1. Summary of tests of validity and reliability of a discrete-choice experiment identified in the literature.

<table>
<thead>
<tr>
<th>Category</th>
<th>Definition</th>
<th>Test</th>
<th>Study design adaptions</th>
<th>Application example</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Measurement validity (8 tests)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Face/content validity</td>
<td>Whether the choice task accounts for important preference attributes and whether results are consistent with a priori preference expectations</td>
<td>Direction of preferences: Compare study results with previously formulated hypothesis</td>
<td>Create different instrument versions</td>
<td>De Bekker et al. (2010) [36]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Intuition: Compare study results with intuition on preferences</td>
<td></td>
<td>Ryan et al. (2001) [28]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Development process: Evaluate instrument development process</td>
<td></td>
<td>Mühlbacher et al. (2009) [37]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Clinical expertise: Ask clinical experts/patients about relevance of included and/or omitted attributes</td>
<td></td>
<td>Kenny et al. (2003) [38]</td>
</tr>
<tr>
<td>Convergent validity</td>
<td>Whether the results are consistent with other measures that measure the same construct</td>
<td>Across surveys: Conduct multiple preference experiments using different preference-elicitation methods</td>
<td>Add additional choice tasks</td>
<td>Bijlenga et al. (2009) [39]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Within surveys: Add a second preference experiment (different method) to an existing experiment</td>
<td></td>
<td>Hollin et al. (2016) [40]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Research context: Examine study results in context of previously conducted preference studies on the same topic</td>
<td></td>
<td>Janssen et al. (2016) [41]</td>
</tr>
<tr>
<td>External validity</td>
<td>Whether the results can accurately predict preferences and choices outside of the study context</td>
<td>Revealed preferences: Compare study results with revealed preference results (adjust for scaling effects)</td>
<td>Conduct a second experiment</td>
<td>Mark and Swait (2004) [42]</td>
</tr>
<tr>
<td><strong>Measurement reliability (6 tests)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test–retest</td>
<td>Whether the instrument measures preference consistently when administered twice</td>
<td>Across-survey reliability: Repeat the same DCE with the same responders and compare responses</td>
<td>Create different instrument versions</td>
<td>Bijlenga et al. (2009) [39]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Within-survey stability: Include the same choice task, a repeat task, in the experiment twice and compare responses within individuals</td>
<td></td>
<td>Ozdemir et al. (2010) [33]</td>
</tr>
<tr>
<td>Version consistency</td>
<td>Whether different versions of the instrument result in consistent preference estimates</td>
<td>Fixed-choice tasks: Include one or more fixed-choice tasks that are the same across survey versions and compare responses across survey versions</td>
<td>Add additional choice tasks</td>
<td>Janssen et al. (2016) [43]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Small survey changes: Make small changes to survey versions and compare results between versions</td>
<td></td>
<td>Velwijk et al. (2016) [44]</td>
</tr>
<tr>
<td>Holdout prediction</td>
<td>Whether the instrument can predict choices outside the choice model (within the instrument) accurately</td>
<td>Within survey: Include one or more holdout tasks in the instrument that are not used in the choice model estimation. Use the choice model to predict choices for these holdout tasks and compare with observed choices</td>
<td>Create different instrument versions</td>
<td>De Bekker et al. (2010) [36]</td>
</tr>
<tr>
<td>Choice validity (5 tests)</td>
<td></td>
<td>Across survey: Include survey version that is not included in the estimation of the choice model. Use the choice model to predict choices for these holdout tasks and compare with observed choices</td>
<td></td>
<td>Rockers et al. (2012) [45]</td>
</tr>
<tr>
<td>Monotonicity</td>
<td>Whether participants do choose ‘worse’ choice profiles over ‘better’ profiles</td>
<td>Within-set dominated pairs: Test whether participants choose the choice profile from a choice task that has better attribute levels than the other choice task by including a dominated choice task in the design</td>
<td>Add additional choice tasks</td>
<td>Ozdemir et al. (2010) [33]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td>Ryan et al. (2001) [28]</td>
</tr>
</tbody>
</table>

(Continued)
<table>
<thead>
<tr>
<th>Category</th>
<th>Definition*</th>
<th>Test</th>
<th>Study design adaptations**</th>
<th>Application example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compensatory choices</td>
<td>Whether participants trade between all attributes of choice profiles</td>
<td>Across-set dominated pairs: Test whether participants choose the preferred choice profile if one profile dominates another choice profile across choice tasks by including multiple choice tasks that are related. Include a choice task between treatment A and treatment B, include a choice task between treatment A and C. Ensure that treatment C has better attributes than treatment B</td>
<td>x^c</td>
<td>Ho et al. (2015) [47] Ozdemir et al. (2010) [33]</td>
</tr>
<tr>
<td>Task nonattendance</td>
<td>Whether participants pay attention to the choice tasks</td>
<td>Attribute dominance: Determine whether certain participants always chose the choice profile that has the better level of a particular attribute Attribute nonattendance: Use a multi-step latent class analysis approach to determine whether participants ignore certain attributes in their decision-making</td>
<td></td>
<td>Ho et al. (2015) [47] Lagarde et al. (2013) [30]</td>
</tr>
<tr>
<td>Choice reliability (5 tests)</td>
<td>Transitivity Whether participants make choices that meet the transitive property (if a participant prefers A over B and B over C, they should prefer A over C)</td>
<td>Transitivity test: Include three choice tasks that ask participants to choose between choice profile A and choice profile B, between choice profile B and C, and between choice profile A and C</td>
<td>x^d</td>
<td>Ozdemir et al. (2010) [33]</td>
</tr>
<tr>
<td></td>
<td>Sen’s consistency Whether participants make choices that follow Sen’s contraction and expansion principle</td>
<td>Contraction principle: If a choice set is narrowed, then no unchosen alternatives should be chosen now and no chosen alternatives should be unchosen now. Test by presenting respondents an initial choice task and a follow-on choice task. The follow-on contraction choice task should present fewer choice tasks than the initial choice task Expansion principle: If a choice set is expanded, then no unchosen alternatives from the original set should be chosen now. Test by presenting respondents with an initial choice task and a follow-on choice task. The follow-on expansion choice task should present more choice tasks than the initial choice task</td>
<td>x</td>
<td>Miguel et al. (2005) [32]</td>
</tr>
<tr>
<td></td>
<td>Level recoding Whether participants make choices using the absolute values of numeric attributes</td>
<td>Scope test: Use multiple survey versions with some overlapping numeric attribute levels but with levels that span different ranges to determine whether the overlapping numeric attributes are valued the same between survey versions Simple scope test: Include a numeric attribute for which the differences between the attribute levels are not equal</td>
<td>x</td>
<td>Ho et al. (2015) [47] Johnson et al. (2011) [29]</td>
</tr>
</tbody>
</table>

DCE: discrete-choice experiment. *Definition of tests adapted to specifically apply to DCEs. ^Some tests require adaptions to be made to the study design. Tests that require adaptions are marked in the table. Generally, a test requires the addition of choice tasks, the creation of different survey versions, or a second experiment. **Necessary choice tasks might occur naturally in the experimental design, but the researchers need to check the choice task properties. ^To assess transitivity, at least two, not at least one, additional choice tasks need to be included.
included. While not every attribute that is important to every possible respondent needs to be included, it is crucial that attributes important to a majority of participants are captured [26]. Face and content validity are most often evaluated using one of four tests.

Face validity of results can be examined by setting a priori hypotheses for the preference relation between the attribute levels. For example, it can be expected that a high clinical benefit will be valued more positively than a low benefit, and this hypothesis can be checked by examining the direction of the preference estimates [28,36]. Assessing face validity is a relatively simple test of validity; about 60% of DCEs published between 2009 and 2012 reported on it [22]. Ryan and Gerard [24] suggested that considering how results match ‘intuition’ can assess face validity. For example, it might be expected that patients with experience of an illness value treatment differently than do the general public [24]. We suggest researchers exercise caution when using a priori expectations and intuition; researchers’ assumptions about the preferences of patients might not be accurate, and researchers need to be careful not to impose their own expectations on patients’ preferences.

Content validity can be assessed by examining the instrument development process [27,37], for example, whether experts were consulted and whether pretest or pilot tests were conducted [34]. In addition, clinical experts and study participants can shed light on whether important attributes were omitted and whether the instrument was deemed relevant [24,38].

4.2. Convergent validity (three tests)

Convergent validity measures the extent to which the results from the DCE are consistent with other measures believed to measure the same construct [25]. Preferences elicited for the same treatment from the same population are expected to be similar regardless of the method used to elicit these preferences. Convergent validity is generally assessed using one of three tests.

Comparisons can be drawn between DCE, visual analog scale, and time tradeoff [39] or direct elicitation of willingness to pay (WTP) [52] using different preference-elicitation instruments. Participants should either complete both types of instruments to allow for within-subject comparison or should be randomized to one instrument to minimize sample biases in across-subject comparisons.

Comparisons can also be drawn between different stated-preference studies in the same preference-elicitation instrument. For example, Hollin et al. [53] conducted a best–worst scaling (BWS) experiment to measure treatment preferences. They included a choice-based discrete-choice question after each BWS task that asked whether participants would accept the treatment [53].

If resources are not available to conduct two experiments, researchers can compare preference estimates they obtain against those reported in the literature [41,43]. Care should be taken that preference estimates are appropriately compared, as different studies might use different estimation methods and different attributes. In these cases, it can help to standardize preferences by, for example, calculating WTP estimates [34].

4.3. External validity (one test)

External validity, sometimes referred to as criterion validity, refers to the extent to which the preference results obtained from the instrument can be generalized to situations and people outside the study. In stated-preference studies, it is often described as the capability of a model to accurately predict preferences and choices outside the immediate model context [31]. For DCEs, this is most easily examined by studying the choices people make in the real world and their revealed preferences and comparing those with the preference estimated in the DCE [34]. This can be done by jointly estimating revealed and stated-preference estimates [42]. In the healthcare sector, relatively few studies have examined the external validity of DCEs [22], partially because opportunities to observe preferences through real-life choices in healthcare might be limited [11].

5. Measurement reliability

Measurement reliability captures whether the DCE produces similar results under consistent conditions. Preference reliability includes concepts of test–retest, version consistency, and holdout tasks. It measures whether results can be reproduced or repeated over a given time. Measurement reliability will never be perfect in any situation because of simple variation in responses.

5.1. Test–retest (two tests)

Test–retest compares the consistency of participants’ responses for the same instrument or choice tasks and can be assessed through test–retest reliability and test–retest
stability. Test-retest reliability refers to the extent to which the instrument results in consistent preference estimates over time. This is tested by asking a sample of respondents to complete the same DCE at least two different times. The retest results are then compared with those from the first time the DCE was administered [39]. Test-retest reliability can be examined by calculating the correlation coefficient (r) when using continuous variables and the Kappa coefficient (κ) when using categorical variables [25]. Test-retest reliability can be problematic if a long period is allowed to pass between the exercises, as people’s circumstances and preferences can change. In that case, the study would not be assessing test–retest reliability but rather the temporal stability of preferences [54].

Test–retest stability refers to the extent to which preferences are consistent for the duration of the survey [33]. To assess test–retest stability, the same choice task needs to be presented to participants twice. The proportion of participants that answer this repeat-choice task the same way twice can then be established [33]. A Kappa coefficient can also be used to examine the difference in expected and observed consistency.

Test–retest stability has the advantage that the instrument only needs to be administered once and that preferences are less likely to change over the course of a single instrument than over the period between consecutive survey administrations. However, test–retest stability only measures consistency for a few (at a minimum, one) choice tasks and might therefore give a less accurate view on the reliability of preferences. For example, if the repeat choice task is first presented at the beginning and then at the end of the instrument, learning effects or fatigue might affect the test–retest stability [33]. In addition, the options presented in the repeat task might affect test–retest stability. For example, a particular task might include options to which many people are indifferent. This increases the probability that they choose a choice profile randomly. In this case, test–retest stability can be expected to be low, even if test–retest reliability for the entire instrument could be high.

5.2. Version consistency (two tests)

Version consistency refers to the extent that different versions of the same DCE result in consistent preference estimates. Version consistency is generally tested in two ways. In experimental designs in which different respondents see different combinations of choice tasks, such as blocked or individualized designs, one or more fixed-choice tasks that are the same for each respondent can be included [27]. Responses to this fixed-choice task can then be compared across survey blocks [43] or across random splits in the sample using a simple t-test [55]. The more fixed-choice tasks are included across the survey versions, the more accurately version consistency can be established. However, adding extra tasks to the experimental design decreases statistical efficiency and increases response burden.

Another way to assess version consistency is by making small changes to survey versions, for example by using decision scenarios in which one version includes a product label as one of the attributes and the other version does not [36] or in the framing of the attributes [44]. This approach has the disadvantage that changes to a survey might mean a different preference construct is measured and responses might not be comparable.

5.3. Holdout prediction (two tests)

Holdout prediction refers to the extent to which the choice model predicts choices outside the model, for holdout tasks, but within the DCE [27]. One or more choice tasks, holdout tasks, or fixed tasks that are not part of the experimental design of the DCE can be included in the instrument. Then, preference estimates obtained from the choice tasks that are part of the experimental design can be used to predict the probability that each choice profile in the additional choice task(s) is chosen [56]. These predictions can be compared with the observed proportion of participants that chose each profile [45].

Another method to assess holdout prediction is to use multiple surveys. Preference estimates for one of the survey versions can be estimated and used to predict the choices for the other holdout survey version [46]. Mühlbacher and Johnson [28] noted that it might be easier to predict choices for one task than for another choice task, which could lead to misleading results for this test.

However, these types of holdout prediction tests capture both response reliability and the predictive properties of the preference model used to analyze the results of the remaining questions in the survey and therefore may be of limited use in establishing a clear measure of response reliability.

6. Choice validity

Choice validity examines whether participants engage with the choice task as expected based on assumptions about consumer behavior. It includes concepts such as monotonicity, compensatory preferences, and task nonattendance. These concepts specifically apply to preference-based methods and measure whether choices comply with assumptions about how choices represent preferences, trading, and responsiveness to the choice instrument.

6.1. Monotonicity (two tests)

Monotonicity, also referred to as non-satiation, means that people will not prefer ‘worse’ levels of an attribute to ‘better’ levels. Two types of monotonicity can be tested: within-set monotonicity and across-set monotonicity [27,33].

Within-set monotonicity tests for non-satiation within a choice task and can be assessed by adding a dominant choice task to the instrument design. This means that, for one of the choice profiles in a choice task, all the attribute levels are just as good as the attribute levels of the other profile and at least one attribute level is better [28,32,33]. If uncertainty exists about how participants value attributes; these attributes should be held constant across the choice profiles in the choice task. To satisfy the within-set monotonicity requirement, a participant should not choose the dominated choice profile. In a two-profile case, there is a 50% probability that someone who does not pay attention to the choice tasks will pass the test for within-set monotonicity [33].
Across-set monotonicity tests for non-satiation across choice tasks; it checks whether people choose the preferred choice profile if one profile dominates across choice tasks. For example, participants might choose treatment B over A in one choice task (treatment A is dominated compared with B). If all the attribute levels of treatment C are better than those of B (treatment B is dominated compared with C), they should not choose treatment A over C in another choice task. Choice profiles that can be used to assess cross-set monotonicity can occur spontaneously in an experimental design, especially if a constant opt-out or status quo option is included [33]. However, researchers might need to add one or more choice tasks to assess cross-set monotonicity.

6.2. Compensatory preferences (two tests)

Compensatory preferences refer to the extent that participants trade between all attributes of the choice profiles. DCEs assume that participants consider all attributes in their decision-making and that there is always some improvement in one attribute that compensates for the reduction in another and that the ranges of levels required to accomplish trading are included in the DCE. Non-compensatory preferences can be indicated by either attribute dominance or attribute nonattendance.

Attribute dominance is present if a participant makes choices based on one attribute only or if they choose the choice profile with the best level of a particular attribute for a majority of choice tasks [28,47]. Lack of clarity exists around the cut-off to define attribute dominance; different cut-off points, such as choosing 70% of choice tasks or all choice tasks with the best level for a particular attribute, have been used to define attribute dominance.

When attribute nonattendance is present, respondents make choices while ignoring one or more attributes. Attribute nonattendance can be examined through multi-step latent class analysis approaches [30]. Attribute nonattendance might be a more comprehensive approach to examining non-compensatory preferences in DCE [30], but it is also more complicated to assess. Lagarde [31] argued that attribute dominance can also be conceptualized as attribute nonattendance, since, in attribute dominance, all attributes except for the dominant one are ignored.

6.3. Task nonattendance (one tests)

Task nonattendance refers to the extent that participants pay attention to the choice tasks. It is assumed that participants actively engage in the choice task, but if the task is too cognitively burdensome or not realistic, they might not carefully consider their decisions [27]. Participants who always choose the choice profile in a particular position in the choice task, for example, the choice profile on the right, are likely not paying attention to the choice task [47,48]. The concept of task nonattendance is more easily measured this way in unlabeled DCEs. In a labeled DCE, a particular product label might have always appeared on one side of the choice task. In this case, if the participant always choses the profile in a particular position, it is not clear whether they did not pay attention to the task or whether they based their choice on the label only (a sign of non-compensatory preferences).

7. Choice reliability

Choice reliability examines whether participants make consistent choices that are in accordance with assumptions about consumer choices. This includes the consistency of choice across choice tasks or survey versions and includes concepts such as transitivity, Sen’s expansion and contraction principles, and level recoding [23,33].

7.1. Transitivity (one test)

Transitivity refers to the preference relation between three or more choice profiles. In particular, if treatment A is preferred to B in one choice task, and treatment B is preferred to C in a second choice task, then the transitive property states that treatment A should be preferred to C in a third choice task. To assess transitivity, at least three related choice tasks need to be included in the DCE instrument [33]. The first choice task, the choice between treatments A and B, can be part of the regular experimental design. The other two choice tasks (treatment B vs. C and treatment A vs. C) will most likely need to be added to the instrument outside of the existing experimental design.

7.2. Sen’s consistency (two tests)

Sen’s consistency principles provide a more stringent test of rationality than the traditional monotonicity tests [28]. Sen’s consistency principles [57] consist of both the contraction principle and the expansion principle. Sen’s contraction principle states that if a choice set A is narrowed (to B) and some of the choices from A are still in B, then no unchosen alternatives should be chosen now and no chosen alternatives should be unchosen now. Sen’s expansion principle states that if a choice set A is expanded (to C) and some of the chosen from A are still in C, then no unchosen alternatives from A should be chosen now. To assess the contraction or expansion property, respondents should be presented with an initial and a follow-on choice task that contain different numbers of choice profiles.

For the contraction principle, the initial choice task should contain at least three choice profiles (e.g. treatment A, treatment B, no treatment). The follow-on contraction choice task should present fewer choice profiles than the initial choice task (e.g. treatment A, treatment B). The contraction property is satisfied if a person who chose treatment A in the initial choice task still chooses treatment A in the follow-on contraction choice task [32].

For the expansion principle, the initial choice task should contain at least two choice profiles (e.g. treatment A, treatment B). The follow-on expansion choice task should present more choice profiles than the initial choice task (e.g. treatment A, treatment B, no treatment). The expansion property is satisfied if a person who chose treatment A in the initial choice task does not choose treatment B in the follow-on expansion choice task [32].
7.3. Level recoding (two tests)

Level recoding refers to the extent that participants process the actual numbers presented in continuous or numeric attributes. Recoding may be a strategy for simplifying evaluations of a relatively unfamiliar but important attribute. Instead of interpreting the actual number and the numeric differences between attribute levels, participants might recode the levels to ‘low,’ ‘medium,’ and ‘high’ categories.

Recoding can be assessed using a scope test, which involves creating two survey versions. For the numeric attribute, these two survey versions need to include some of the same levels (overlapping levels) and at least one different level so the level range is different for the two surveys. If preference estimates are the same for overlapping numeric levels between the two survey versions, we can infer that respondents did not recode the levels but rather reacted to the absolute levels. If the preference estimates for the overlapping levels are different, this suggests participants recoded the numeric values as ‘high,’ ‘medium,’ or ‘low’ [47]. The scope test should be used with caution because if the range of the levels for one of the survey versions is too extreme, it could invoke non-compensatory preferences (discussed above).

An approximation of the scope test, a simple scope test, can be done using just one survey version with a large difference and small difference between the levels of the numeric attribute included. If the difference in preference estimates between the large- and the small-level difference are similar, this might indicate that participants recoded the numeric levels as ‘low,’ ‘medium,’ and ‘high’ [29]. However, the simple scope test does not determine with certainty that recoding occurred, because the preference function might not be linear.

8. Study qualities beyond validity and reliability

Validity and reliability are just two issues that affect the study quality of a stated-preference study and a strict focus on only tests of validity and reliability will not guarantee a high-quality preference study. Many recommendations and guidance documents on good research practices exist on the development, design, implementation, and analysis of stated-preference studies [6,9,14,22,24,26,27,40,41,56,58–67]. These recommendations discuss varying aspects of stated-preference studies, ranging from patient relevance to logical soundness, that researchers should consider when conducting and evaluating stated-preference studies. In many of these recommendations, validity and reliability play a role, but are not the only study qualities that are discussed.

Stakeholders’ views on the importance of different aspects of study quality, including the importance of ensuring validity and reliability, are largely unknown. It is not clear which study qualities stakeholders believe should be prioritized to advance the use of stated-preference studies. While it might be desirable to explore certain study qualities in terms of the scientific advancement of stated-preference studies, these qualities might not currently be actionable.

We engaged a diverse group of stakeholders to discuss and prioritize important characteristics of preference studies and where validity and reliability fell within this prioritization. Participants (n = 29) from industry, regulatory agencies, funding agencies, patient advocacy organizations, and academic institutions with experience conducting or evaluating stated-preference studies were invited to participate in an all-day workshop on stated-preference methods. As part of the workshop, we asked participants to form small groups and list the important characteristics of a stated-preference study. Then, each participant allocated 12 votes across the identified characteristics based on what they believed were the most desirable aspects for high scientific quality. Participants could allocate one vote to 12 characteristics, allocate 12 votes to one characteristic, or allocate their votes in any other combination. Next, participants repeated the voting process, but this time they allocated 12 votes to the characteristics that they thought were most actionable. After the workshop was completed, the study team categorized the characteristics identified by participants into aspects of a stated-preference study. These aspects were ranked for desirability and actionability according to the number of votes received from the workshop participants.

Through this engagement process, eight aspects of stated-preference studies that determined study quality were identified. These aspects are presented and described in Table 2. Aspects identified included respondents’ understanding, appropriate research question, diverse samples, transparency of methods, patient/population centered, stakeholder relevance, internal validity, and external validity.

Understanding/interpretation by respondents was prioritized as the most desirable and actionable aspect. Internal validity was ranked as the seventh most desirable aspect and the sixth most actionable aspect of a stated-preference study. External validity was prioritized as the fourth most desirable but the eighth most actionable aspect. These results suggest that stated-preference stakeholders might be particularly interested in the conceptualization of quality as a process measure [31] rather than solely as an outcomes measure. Further work on the implementation of good research practices for instrument development to ensure process validity will be essential [63].

9. Expert commentary

To be able to conduct many of the discussed validity tests, adaptations to the study design of a DCE are necessary (Table 1). These adaptations need to be taken into account before the study commences and might also complicate data analysis. Design adaptations might increase study costs, decrease statistical efficiency, or increase response burden. Face validity, content validity, attribute dominance, and attribute nonattendance do not require special study design considerations so can be relatively easily incorporated into almost any DCE if the entire study process, including instrument development, is transparently reported. While attribute nonattendance does not require special study design considerations, it does require the use of a relatively complex analytical process. In addition, external validity does not require changes to the study design, but revealed preferences might not exist and/or might require...
complex statistical analyses. Therefore, these tests might be more difficult to conduct.

To be able to assess monotonicity, transitivity, Sen’s consistency principles, and test–retest stability, at least one additional choice task needs to be added to the study design. To be able to assess level recoding, different instrument versions need to be administered to different participants. To be able to assess test–retest reliability, the DCE needs to be administered twice to the same group of participants. Convergent validity can be tested by conducting a second choice experiment either as part of the DCE instrument or outside of it. Version consistency and holdout predictions can be assessed either by adding a fixed-choice or holdout task or by administering multiple survey versions to different participants.

In cases where validity tests can be performed using multiple study design adaptations, the type of adaptation that requires more resources generally presents a more stringent test. For example, to assess level recoding, a scope test requires that different survey versions are administered to different participants. The simple scope test only requires the adaption of the numeric attribute levels, but this test does not conclusively show whether recoding occurred or whether preferences are not linear. Generally, more stringent tests require larger adaptations to the study design [28]. Within-set monotonicity is regarded as a weak test of choice validity because it can be relatively easily passed by chance [33], but assessing within-set monotonicity only requires one additional choice task. Transitivity and Sen’s consistency theory are more stringent tests of choice reliability but require larger changes to study design. Assessing transitivity requires the addition of at least two choice tasks. Assessing Sen’s consistency principle requires the addition of at least one choice task that does not follow the format of the other choice tasks, which might add complexity to the instrument.

Researchers need to be careful in the interpretation of tests of validity and reliability and the conclusions reached based on these tests. Apparent violations of face validity may reflect incorrect assumptions about patient preferences. Furthermore, failure of the dominance test might result from lack of understanding on the part of the researcher as to the interpretation of attribute levels by participants. In addition, the estimation of choice data assumes an element of randomness in the choices. This randomness needs to be considered when conducting and interpreting validity or reliability tests – what may appear to be an inconsistent response may just be noise and not a violation of the rules.

Observed violations in the validity and reliability of tests can either be interpreted as an issue with the DCE or as an issue with the respondents. It is important to keep in mind that observed violations do not necessarily mean participants did not understand the tasks, did not use logic in their choices, or were inattentive to the choice tasks. Instead, unexpected results might signal problems in the design and implementation of the DCE [68,69]. It might signal that the researcher did not ask the question appropriately, did not understand the responses appropriately, or imposed their own assumptions on what preferences should be. Furthermore, different theoretical assumptions [50,70–73] might lead to different expectations about the way participants will complete DCE choices.

When interpreting preference results, researchers should start with the assumption that measured responses reflect participants’ preferences. Violations of tests of validity and reliability or other unexpected results should not be interpreted as a failure of participants. Rather they should be taken as an opportunity for the researcher to reflect on their preference instrument, assumptions, and a priori expectations.

### 10. Five-year review

Unanswered questions surrounding the validity and reliability of DCEs will need to be addressed. While a variety of tests exist to assess the validity of a DCE, important questions remain. It is unclear what to do with participants who fail tests for validity or reliability [28]. Excluding participants who fail the test for choice validity or reliability might increase internal validity but will decrease the external validity of results. Deleting responses can increase sample selection bias and decrease the statistical efficiency and power of the estimated choice models [69]. In addition, participants who fail these tests might have made their choices according to their actual preferences. Follow-up
questions or debriefing interviews could shed light on the way participants answer choice tasks [32]. It remains unclear exactly when a test of validity or reliability is met. This is especially difficult for tests that require some qualitative or subjective assessment such as face or content validity. However, ambiguous requirements for more quantitative tests also exist. For example, the acceptable frequency of choosing the dominated treatment profile, or the failure rates of rationality tests or test–retest stability in a study, are not well established. It is likely that ambiguity on when a test is met will remain even if more research is conducted. Such ambiguity can be incorporated into formal tests with, for example, the establishment of ‘grey’ zones in which conclusions on validity or reliability remain inconclusive [74]. To be able to answer questions of ambiguity, more research on theoretical properties of DCEs and stated-preference methods is needed. A review of theoretical contributions to the questions of validity and reliability would be needed to identify gaps in the literature.

It is also what set of tests needs to be met for a DCE study to be considered valid. It is unlikely that one study can incorporate each test of validity or reliability, especially because many tests require an adaptation to the study design. A standard set of tests to incorporate is unlikely to be applicable to every DCE. Different aspects of study quality, including issues of process validity [31] or the aspects identified by stakeholders presented in this study, might be more important to consider for certain purposes. Therefore, further work should examine which tests and study aspects might be most important to include given a particular research question and population.

11. Conclusion

The complexity of DCEs means most studies will neither include, nor meet all tests of validity and reliability identified in this review. More studies are needed to investigate the determinants of validity for a DCE; this work might include studies on study quality and good research practices that go beyond the concepts of validity and reliability as outcome measures. Given unanswered questions regarding tests for validity and reliability, researchers need to remain cautious when applying these tests. Personal expectations and assumptions regarding the outcomes of these tests should be carefully examined so as not to lead to incorrect conclusions. While limitations in establishing the reliability and validity of a DCE remain, this review can help researchers gain insights into the available types of tests for validity and reliability and how to incorporate these tests into their studies.

Key issues

- The recent endorsement of stated-preference methods, including discrete-choice experiments (DCEs), by regulatory and health technology assessment (HTA) agencies has placed a greater focus on demonstrating the validity and reliability of preference results.
- From the published literature, we developed a framework to conceptualize tests of validity and reliability for DCEs. This conceptual model spans 24 tests and includes four conceptual domains: measurement validity, measurement reliability, choice validity, and choice reliability.
- Validity and reliability are just two important aspects of the quality of a stated-preference study. Other important aspects, as identified by stakeholders, include respondents’ understanding, appropriate research question, diverse samples, transparency of methods, patient/population centered, stakeholder relevance, internal validity, and external validity.
- In the next 5 years, unanswered questions regarding the validity and reliability of DCEs need to be addressed. These questions include treatment of participants who fail tests for validity and reliability, selecting cut-off points to determine when a test of validity and reliability is met, and providing guidance on when a stated-preference study can be considered valid.
- Caution needs to be taken when assessing reliability and validity for stated-preference studies as many questions regarding the application of these tests remain. Furthermore, personal expectations and assumptions regarding the outcomes of these tests should be carefully examined so as not to bias conclusions that are based on these tests.

Acknowledgments

The authors sincerely thank the Johns Hopkins Institute for Clinical and Translational Research (ICTR) Community Research Advisory Council (C-RAC) and members of the Diabetes Action Board (DAB) for their valuable contributions and engagement in the research study.

Funding

This work was supported by a Patient-Centered Outcomes Research Institute (PCORI) Methods Award (ME-1303-5946) and by the Center for Excellence in Regulatory Science and Innovation (CERSI) (1U01FD004977-01).

Declaration of interest

DA Marshall is funded through a Canada Research Chair, Health Systems and Services Research and the Arthur J.E. Child Chair Rheumatology Outcomes Research. A8 Hauber is an employee of RTI Health Solutions. The funders had no role in the design and conduct of the study, interpretation of the data, or preparation of the manuscript. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed. Peer reviewers on this manuscript have no relevant financial or other relationships to disclose.

References

Papers of special note have been highlighted as either of interest (i) or of considerable interest (ii) to readers.


10. The Framework Report discusses considerations for the use of patient preference information in the regulatory assessment, reimbursement, marketing, and shared decision-making. It specifically calls for the need to study what determines the validity and reliability of stated-preference studies.


16. Guidance put forth by the US FDA on patient preference information (PPI) that may be used by FDA staff in decision-making on medical devices. The guidance presents 11 study qualities to consider when conducting patient preference studies.


39. Bijlenga D, Birnie E, Bonsel GJ. Feasibility, reliability, and validity of three health-state valuation methods using multiple-outcome